NAVAL HEALTH RESEARCH CENTER

MODELING RUN TEST VALIDITY: A META-ANALYTIC APPROACH

R. R. Vickers, Jr.

Report No. 02-27

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CA 92186-5122

BUREAU OF MEDICINE AND SURGERY (MED-02)
2300 E ST. NW
WASHINGTON , DC 20372-5300

Modeling Run Test Validity:
A Meta-Analytic Approach

Ross R. Vickers, Jr.

Human Performance Program
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

email: Vickers@nhrc.navy.mil

Report Number 02-27, supported by the U.S. Army Medical Research and Materiel Command, Ft. Detrick, Frederick, MD, under Work Unit No. 60109. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

This research has been conducted in compliance with all applicable Federal Regulations governing the protection of human subjects. No human subjects were directly involved in this research.

Approved for public release; distribution unlimited.

SUMMARY

Background

Military physical fitness tests (PFTs) commonly include a distance run to estimate aerobic fitness. This use of run tests is justifiable because performance is related to laboratory measurements of maximal oxygen uptake (VO_{2max}). VO_{2max} is the accepted reference standard for measuring aerobic capacity.

Objective

This review tested the hypothesis that endurance run tests are equally valid for different types of people.

Approach

Literature searches identified 133 studies relating laboratory VO_{2max} measurements to performance on run tests. These studies involved 1 or more endurance runs. An endurance run was defined as one that was ≥2 km in distance or ≥12 min in duration. Data on validity, population attributes, and research methods were coded. Hedges and Olkin's meta-analysis procedures were used develop a mathematical model to predict run test validity.

Results

Average validity was moderately high ($r = .75$), but could vary between $r = .52$ and $r = .84$ in different test situations. Run test validity was not related to the age, gender, fitness level, or running experience of the people who were tested. Validity was related to the type of run test (T for "type," 1 = fixed-distance; 2 = fixed-time) and the sample variability of VO_{2max} ($SDVO_{2max}$ for standard deviation of VO_{2max}). The equation to predict validity based on these two variables was $z_{UF}' = .059 + (.115*SDVO_{2max}) + (.197*T)$. The equation coefficients indicate that validity was higher in samples with greater VO_{2max} variation and when a fixed-time run test was used. Publication bias, low power of the studies reviewed, and other potential sources of bias had little effect on the findings.

Conclusions

Department of Defense PFT run test components are equally valid regardless of the age, gender, or fitness level of the people tested. Fixed-time tests (e.g., a 12-min run) might be preferable to fixed-distance tests. However, this point is debatable because few studies have directly compared the 2 types of test. Further study would be needed to determine whether switching to fixed-time tests could increase the validity of PFT assessments of aerobic fitness.

Introduction

Endurance running performance is strongly related to maximal oxygen uptake capacity (VO_{2max}; Baumgartner & Jackson, 1982; Knapik, 1989; Safrit, Hooper, Ehlert, Costa, & Patterson, 1988; Vickers, 2001a, 2001b). Early reviewers placed the association between r = .60 (Katch & Henry, 1972) and r = .66[1] (Safrit et al., 1988).

The average validity[2] of endurance runs justifies the use of running performance as a method of estimating VO_{2max} in many testing situations. However, for at least 20 years, physical fitness assessment experts have cautioned that the validity of running performance as a basis for estimating VO_{2max} may depend on the age, gender, running experience, or other attributes of the people tested (Baumgartner & Jackson, 1982). Average validity, therefore, may be misleading.

Systematic quantitative investigation of factors that affect run test validity has been limited to a review by Safrit et al. (1988). Those authors concluded that validity coefficients should not be generalized. Their conclusion was based on the variation in run test validity coefficients within different groups. The evidence indicated that this variation was too great to be dismissed as chance when different samples of men were compared. The evidence was less clear for boys and girls because data were more limited for these groups. The evidence that was available suggested that validity would not generalize within these groups either. Women were the only group with coefficients that were stable enough to generalize validity to new samples with confidence.

Safrit et al.'s (1988) validity generalization analyses focused the amount of variation in validity coefficients within populations. The search for variables that could explain the observed variation was limited to the effects of test reliability. A broader search for explanatory variables might change the conclusions that would be drawn from the evidence. For example, Vickers (2001a, 2001b) found that the length of the run test affected validity. He also found that validity was lower for fixed-distance runs (r = .72) than for fixed-time runs (r = .80). Both factors could account for some of the within-population variance noted by Safrit et al. (1988). If these factors account for enough variance, the residual variation might be no greater than expected by chance. If so, the evidence would support the

[1] The author computed this value from the correlations reported in Safrit et al. (1988). Those authors only reported a value of r = .741 obtained after correcting for unreliability of the measures.

[2] Validity refers to the appropriateness of a proposed interpretation of test scores (American Psychological Association, 1985). A given test may have more than one interpretation. Each interpretation can be valid if there is empirical evidence to support that interpretation. In this report, run test validity refers specifically to run test performance as an indicator of maximal aerobic capacity.

view that validity generalizes within populations, provided the run test is long enough and the type of test is specified.

Safrit et al. (1988) regarded their results as a framework for further study. That framework is the starting point for the present investigation. The general objective is to provide a more detailed model of run test validity. The model development process addresses several general questions. Is validity related to the age, gender, fitness level, or running experience of the study participants? How much within-population variation can be accounted for by methodological factors? Do population attributes and methodological factors combined account for enough variance to conclude that validity generalizes across samples? Answers to these questions would help users of run tests select the best available tests for their purposes and interpret the results of those tests correctly.

The answers to the questions posed above are not self-evident. For example, age or gender might appear certain to affect run test validity, but the plausibility of these effects is not proof that they are present. In fact, Safrit et al.'s (1988) data suggest that these factors have little effect on validity. Those authors reported average validity coefficients of $r = .69$ for men, $r = .58$ for women, $r = .70$ for boys, and $r = .64$ for girls.[3] Analysis of their data indicates that gender affects average validity, but that there is no effect of age and no age by gender interaction.[4] Thus, their data provided empirical support for only 1 of the 3 hypotheses about age and gender effects.

Using the Safrit et al. (1988) framework as a point of departure, this review extends the analysis of run test validity:

A. More evidence is considered. Safrit et al. (1988) covered 34 studies. This review covers 133 studies reporting on 166 samples.

B. Each study contributes a single validity coefficient. Validity coefficients are averaged when more than 1 correlation is reported for a sample.[5] Safrit et al. (1988) treated each validity coefficient separately. Their approach could yield misleading statistical tests because the observations being analyzed are not independent (Becker & Schram, 1994; Raudenbush, 1994). Developments in meta-analytic practice

[3] The author computed the values from data reported in Safrit et al. (1988). Safrit et al. (1988) actually reported values of $r = .77$ for men, $r = .64$ for women, $r = .78$ for boys, and $r = .71$ for girls. Those coefficients included corrections for reliability. The reliability data reported by Safrit et al. (1988) were used to obtain raw correlations by reversing the correction. The raw correlations are of interest here because they represent what would actually be observed in a validation study.

[4] Significance determined from analyses conducted by the present author using Hedges and Olkin's (1985) procedures for a 2-way analysis of variance with gender and age (<16, >16) classification variables.

[5] An exception to this procedure was made for 3 of 133 studies. See Methods.

since 1988 have made it common practice to minimize this problem by averaging.

C. Coverage is limited to fixed-distance runs ≥2 km and fixed-time runs ≥12 min. Run test validity increases with test length for short runs, but validity is constant for runs that meet either of these criteria (Vickers, 2001a, 2001b). Safrit et al. (1988) included fixed-distance runs ≥1 mile and fixed-time runs ≥9 min. The inclusion criterion used in this review focused the review on run tests with maximal validity.

D. Test type is considered as a source of variance. Validity is higher for fixed-time runs than for fixed-distance runs (Vickers, 2001a, 2001b). Test type differences are a plausible source of some of the within-population variation identified in Safrit et al.'s (1988) review.

E. Age and gender effects on validity are quantified. Safrit et al. (1988) did not formally evaluate the magnitude of these effects.

F. Fitness level and running experience are evaluated as influences on validity. Baumgartner and Jackson (1982) suggested that these attributes could affect run test validity. To date, these suggestions have not been analyzed formally.

G. The variability of VO_{2max} within a sample is evaluated as an influence on the validity coefficient observed in that sample. Safrit et al. (1988) did not correct for restriction of range, a statistical artifact that can affect validity generalization estimates (Hunter & Schmidt, 1990).

H. A multivariate model is developed to account for study-to-study differences in validity. This model reduces the risk of obtaining biased estimates of effects. Effects can be biased if important influences on run test validity are omitted from the statistical model. Bias will occur if the omitted variables correlate with one or more predictors in the model (cf., James, Mulaik, & Brett, 1982, pp. 71-80).[6]

I. Hedges and Olkin's (1985) meta-analysis procedures are used. Safrit et al. (1988) employed the Hunter, Schmidt, and Jackson (1982) validity generalization approach. The two approaches yield similar statistical inferences (Schmidt & Hunter, 1999). However, the Hedges and Olkins (1985) approach can be adapted to contrast fixed-effect (FE) and random-effect (RE) models (Hedges & Vevea, 1998). The RE approach has been recommended

[6] For example, consider a possible analysis of age effects. Vickers (2001a,b) found that runs of 1 mile or less are less valid than longer runs. Suppose that short runs are used more often in studies of children than in studies of adults. Age and run distance will be correlated. Estimated effects of age on validity will be biased in analyses that do not control for this correlation. James et al. (1982) referred to this potential problem as omitted variable bias. The solution is to identify potential sources of bias (e.g., run distance) and include an indicator for each source in the analysis if possible (James et al., 1982).

by the National Research Council (1992) to avoid overstating the precision of meta-analytic findings.

J. Sensitivity tests are conducted. Meta-analysis is a complex process involving a number of decisions, each of which may affect the findings (Wanous, Sullivan, & Malinak, 1989). Biases can be introduced from various sources, including publication bias (Rosenthal, 1984), outlier data points (Hedges, 1987; Hedges & Olkin, 1985), and the treatment of studies with small sample sizes (Kraemer, Gardner, Brooks, & Yesavage, 1998). The National Research Council (1992) recommended the routine use of sensitivity tests to evaluate the effects be a routine part of meta-analyses to ensure that results are interpreted properly.

These elaborations on Safrit et al.'s (1988) inaugural efforts should further the understanding of run test validity. If one or more of the factors explored here affects run test validity, the results will help clarify the boundaries of validity generalizations.

Methods

Literature Search

The literature search was conducted in two phases (Vickers, 2001a, 2001b). Elements of the first phase were:

A. An initial list of articles was constructed from prior reviews (Baumgartner & Jackson, 1982; Knapik, 1989; Safrit et al.).

B. A keyword search was conducted for the MEDLINE™, PsychLit, and Discus databases. The primary search terms were "maximal oxygen uptake" paired with "run time" or "running" (Vickers, 2001a).

C. An ancestry search (Rosenthal, 1984; White, 1994) was conducted. The full reference lists in the articles identified in steps A and B were reviewed by title. Articles that mentioned running and VO_{2max} were added to the list of potential data sources. The bodies of the articles were reviewed to identify citations in sections that summarized evidence relating physiological variables to running performance. Those articles were added to the list of potential data sources.

D. A year-by-year search was conducted by examining the tables of contents for the *Journal of Sports Medicine and Physical Fitness*, *Medicine and Science in Sports and Exercise*, *European Journal of Applied Physiology*, and *Research Quarterly for Sports and Exercise*.

E. The PubMed® database was searched to identify papers published while the preceding steps were being completed.

F. The "related articles" option in PubMed was examined for each new article found. This step extended the earlier ancestry search.

G. The catalogues at San Diego State University, the University of California, San Diego, and the Naval Health Research Center were searched to identify unpublished master's theses, doctoral dissertations, and technical reports.

The second phase of the search repeated steps B through D and G of the first search (Vickers, 2001b). In this phase, the term "threshold" was substituted for the oxygen uptake terms used in the initial search. The objective in the second search was to determine whether the set of studies identified in the first review could be expanded by locating research that focused primarily on constructs such as anaerobic threshold, ventilatory threshold, or onset of blood lactate accumulation (Vickers, 2001b). Steps E and F were not repeated because this portion of the review was completed relatively quickly.

The two searches identified 133 studies published between 1964 and 2000 with at least one run test that met the inclusion criteria for this review. The median publication date was 1984. Most (83.4%) studies had been published as journal articles. The other studies included master's theses, doctoral dissertations, technical reports, and papers presented at meetings or symposia. The 133 studies presented results for 166 samples. The median sample size was $n = 23$ (range = 7 - 866).

Study Attributes

The following information was extracted from each article when available:

A. *Year*. The year of publication or presentation of the findings was recorded.
B. *Publication status*. This variable was coded as journal article or nonjournal publication.
C. *Sample size*. This variable was the number of participants who contributed data to each correlation coefficient included in this meta-analysis. In studies that reported more than one coefficient, missing data meant that sample size varied from one coefficient to another. In these cases, a separate sample size was coded for each coefficient. When the computation of an average effect size was necessary, the average was computed using the actual sample size for each coefficient. The smallest sample size for any coefficient in the set used to compute the average then was used as the sample size for the computations reported in this review.
D. *Age*. The average age for the participants was recorded when reported. If a study reported only an age range (e.g., 8- to 10-year-olds) or an implied age range (e.g., 6th to 9th grade students), the midpoint of the range was used to estimate sample age. The midpoint is a reasonable estimate if the age distribution is approximately uniform over the range. Age was not estimated for samples described as "college-aged" or "college students" because of the risk of erring by several

years. The age range for college students is wide and seemed likely to vary from one institution to another. The distribution almost certainly is not uniform over the range, even if the range were known. Thus, knowing that a sample was drawn from a college population provided too little information to estimate average age with confidence.

E. *Gender*. This variable was coded male, female, or combined male and female. Analyses assessing gender effects compared samples of males with samples of females. The samples that combined males and females were omitted from the analyses assessing gender effects.

F. *Running experience*. The basic categories for this variable were coded as untrained or endurance athlete. The variable also included codes for "other athlete" (i.e., any athlete who did not train for endurance running), and mixtures (i.e., samples with some combination of the other 3 categories). Analyses to assess the effects of running experience contrasted just the first two groups. Samples characterized as recreational runners were placed in the untrained category. The experience concept was intended to differentiate people who had run competitively from people who had not. Competitive runners were expected to know their limitations with greater precision and to understand the running strategy that would get the best performance from their abilities. Recreational runners might not have tested their upper limits in their running and would be expected to have less experience at maximizing their performance. Marathon runners were the exception to this rule. Given the amount of training typically involved in preparing for a marathon, all runners who participated at this distance were classified as endurance athletes.

G. *Method of measuring maximum oxygen uptake* (VO_{2max} Protocol). Oxygen uptake measurement protocols varied from study to study. The differences were coded in terms of the major elements of the protocol. These elements were the exercise performed and the methods used to measure oxygen uptake. The protocols were coded as continuous treadmill with open spirometry (CTO), continuous treadmill with Douglas bag (CTB), intermittent treadmill (IT), and all other procedures (Other). Analysis of specific protocol attributes (e.g., frequency and magnitude of increases in treadmill speed and/or slope) was beyond the scope of this review.

H. *Maximal aerobic capacity* (VO_{2max}). The average maximal oxygen uptake in $ml \cdot kg^{-1} \cdot min^{-1}$ for the sample was recorded when reported. A few studies reported data for individuals, but did not report the average. The author computed the average in these cases. VO_{2max} and VO_{2peak} were treated as equivalent measures.

I. *Standard deviation of VO_{2max}* ($SDVO_{2max}$). The sample standard deviation for VO_{2max} was recorded when reported or could be computed from raw data reported in the study. Care was taken to distinguish between studies that reported the standard error of the mean (SEMs) and those that reported the standard deviation. When the SEM was reported, the standard deviation was computed by multiplying the SEM by the square root of the sample size.[7]

J. *Distance*. This variable was a constant for all participants when the run covered a fixed distance (e.g., 5 km). The average distance was recorded for fixed-time tests (e.g., 12 min).

K. *Time*. This variable was the set duration for fixed-time tests. Average time was recorded for fixed-distance tests.

L. *Validity Coefficients*. The validity coefficients were Pearson product moment correlations of VO_{2max} with running performance. The database included one coefficient for each of the 166 samples covered in the review. Several steps were taken to ensure that coefficients were comparable across samples. First, the sign of the correlation was reversed when performance was measured as the time taken to run a fixed distance. The typical correlation in this case initially was negative. Other performance criteria, such as the distance covered in a fixed-time run or the average velocity in a fixed-distance run, produced positive correlations. A positive correlation seemed more natural when making generic statements about the association (i.e., high VO_{2max} = good performance). After reversal, the database included only one negative correlation for a run that met the current endurance test criteria. This negative relationship was a case in which high VO_{2max} actually was associated with poorer performance. The finding presumably was a chance result. The great preponderance of evidence, therefore, paired high VO_{2max} values with better than average running performance.

Validity coefficients were averaged when more than one relevant correlation was reported for a sample. In earlier reviews relating validity to the distance/duration of the run, Vickers (2001a, 2001b) treated each run test as a separate case. This approach made it possible to analyze the relationship between validity and distance/time in detail in those studies. However, the results of those earlier analyses were used to limit the present review to run tests that shared a common validity within test type (i.e., fixed-distance or fixed-time). Given this restriction, each qualifying coefficient reported for a given sample estimated that common validity. Under these conditions, the average was the best estimate of run test validity in that sample.

[7] Kearney and Byrnes (1974) reported a small $SDVO_{2max}$. A figure in the paper showed a wide range of values. The present author estimated individual data values from the figure. Analysis of the estimates indicated the reported $SDVO_{2max}$ really was the SEM. The SEM adjustment therefore was applied to the reported $SDVO_{2max}$.

An exception to the averaging procedures was made when a study reported both fixed-distance and fixed-time tests for a given sample. Vickers (2001a,b) showed that average validity differed for the two types of test. This difference made averaging across test types inappropriate because two distinct parameters apparently were involved. In fact, test type was one variable that was to be examined in this review. Thus, fixed-distance and fixed-time tests reported for a given sample were kept separate. This decision meant that 3 samples were represented in the data set by a coefficient for one or more fixed-distance tests and by a separate coefficient for a fixed-time test.[8]

Analysis Procedures

Rosenthal and DiMatteo (2001) captured the spirit of meta-analytic data analysis in two observations. "Meta-analysis is not inherently different from primary data analysis; it requires the same basic tools, thought processes, and cautions" (Rosenthal & DiMatteo, 2001, p. 78). "The best quality scientific exploration is often one that poses unadorned, straightforward questions and uses simple statistical techniques for analysis" (Rosenthal & DiMatteo, 2001, p. 68).

The preceding quotes have been noted because meta-analysis can appear to be more complex than primary data analysis. Any meta-analysis involves a number of decision points (Wanous et al., 1989). Also, effect sizes are analyzed rather than raw data. Despite these differences, meta-analysis basically involves a set of standard issues and concerns analogous to those encountered in the design, conduct, and analysis of a primary study. The decision processes are analogous to common research design decisions, such as the design of a sampling frame and treatment of outlier data points. The statistical analyses rely on the same computational procedures used to compute descriptive statistics, analysis of variance (ANOVA), and regression in the analysis of primary data.

Standard computerized data analysis packages can be used to conduct meta-analyses (Hedges & Olkin, 1985). In the present case, the SPSS-PC (SPSS, Inc., 1998a, 1998b) program was used to perform the following analytic steps:

A. Olkin and Pratt's (1958) correction for sample bias in the estimated correlations was applied. Hedges and Olkin (1985) noted

[8] Knowlton and Gifford (1972) included fixed-distance and fixed-time tests that qualified for this review. However, a correlation was reported only for the fixed-time test. An estimate of the fixed-distance correlation could have been derived from the reported proportion of variance explained. However, a decision to use only the reported correlation was made in the initial review of the article (Vickers, 2001b). That decision was carried over to the present paper to base analyses on a consistent set of data.

that this correction is most important when $0.4 < r < 0.6$ and sample size is small (e.g., $n < 15$). The average correlation and median sample size in the present analyses were slightly higher than these figures, but the correction was retained to keep the analysis procedure comparable to those used by Vickers (2001a, 2001b).

B. Fisher's *r*-to-*z* transformation was applied to normalize the distribution of correlations (Hays, 1963). The data points analyzed, therefore, are labeled $z_{UF(i)}$ as a reminder that they are **u**nbiased, **F**isher-transformed estimates of the population correlations for a given sample, denoted by the "*i*" in the subscript.

C. The $z_{UF(i)}$ were averaged when more than one correlation was reported for a sample. Vickers (2001a, 2001b) did not average because there was reason to believe that runs covering different distances would have different validity coefficients. However, those analyses showed that the endurance runs considered here shared a common average correlation. Averaging, therefore, was appropriate because the correlations were believed to estimate a single population parameter. Averaging avoided statistical problems associated with nonindependence of observations (Becker & Schram, 1994; Steiger, 1980). Four exceptions were made with regard to averaging. Two correlations reported by Iwaoka, Hatta, Atomi, and Miyashita (1988) differed so widely that an average seemed likely to misrepresent the data. This study ($N = 10$) was dropped from the analysis. Three studies reported both a fixed-time test and a fixed-distance test (McNaughton, Hall, & Cooley, 1998; O'Donnell, Smith, O'Donnell, & Stacy 1984; O'Gorman, Hunter, McDonnacha, & Kirwan, 2000). In these instances ($\Sigma N = 89$), the correlations were treated as separate cases to simplify the examination of test type as an influence on validity. After averaging, the data set consisted of 169 effect size estimates from 166 samples of subjects who participated in 133 studies. Ten of 5,767 participants had been dropped. The 89 participants who completed both fixed-time and fixed-distance tests comprised 1.5% of the total participants who contributed data to the analyses. This slight departure from independence of the correlation coefficients was accepted to increase statistical power when testing of differences between types of run test. The risk that the lack of independence would seriously distort the analyses appeared minimal.

D. Each correlation was compared with a predicted value (i.e., $z_{UF(i)} - z_{UF(i)}'$). The predicted values were familiar elements of standard analysis procedures. For example, the predicted values in one ANOVA were the sample-size-weighted means for 4 different methods of measuring VO_{2max}. The predicted values in another analysis were determined from the regression of $z_{UF(i)}$ on the logarithm of distance.

E. The difference between the observed and predicted values was standardized. This was accomplished by dividing $z_{UF(i)} - z_{UF}'$ by the standard deviation for the transformed correlation (i.e., $1/\sqrt{[N_i - 3]}$).

F. The standardized value for the difference was squared to produce a χ^2 with 1 degree of freedom (*df*) (Hays, 1963).

G. The χ^2 values for all correlations in the analysis were summed to produce an overall χ^2 that was the summary fit statistic for the model.

H. The χ^2 values for competing models were compared to determine which model best accounted for the observed variation in the correlations.

This somewhat detailed summary shows that meta-analytic computations revolve around the size of differences between observed values and the predicted values estimated from different statistical models. These computations are directly analogous to the deviations and/or residuals computed for descriptive statistics, ANOVA, or regression analyses of raw data. Also, the statistics used for comparing alternative models are comparable to using incremental variance explained in stepwise primary data analyses.

Meta-analysts must choose between FE and RE models (Hedges & Olkin, 1985; Hedges & Vevea, 1998; Raudenbush, 1994). An RE model was the intended end point of this review. This end point was selected to provide a suitable basis for drawing the widest possible inferences about run test validity. Developing a FE model would limit the inferences to studies similar to those in the review (cf., Hedges & Vevea, 1998).

This review adopted a two-stage analysis approach even though an RE model was the intended end point. FE models were used for initial screening of effects. FE models have a smaller error variance term than RE models (Becker & Schram, 1994; Erez, Bloom & Wells, 1996; Hedges & Vevea, 1998). FE significance tests will be lenient when an RE model is appropriate. Lenient tests in the initial analyses reduced the risk that one or more influences on run test validity would be prematurely eliminated from consideration for the multivariate validity model. Also, an FE model is the point of departure for estimating the additional RE variance in the Hedges and Vevea (1998) approach used here. Thus, significant FE effects were examined further in the second stage of the analyses by computing an RE model following procedures described by Hedges and Vevea (1998). Comparing results from the FE and RE models provided a sensitivity test for model choice as recommended by the National Research Council (1992).

Analyses were conducted with the general linear model (GLM) and linear regression procedures in SPSS-PC (SPSS, Inc., 1998a, 1998b). The weighted least squares option in each procedure was used with (n - 3) as the weight for each transformed correlation. Using this weighting option, the sums of squares reported in the analysis results are χ^2 values (cf., Hedges & Olkin, 1985, pp. 235-241). The GLM procedure was used when bivariate analyses involved discrete groups (e.g., males and females) and for the construction of multivariate models. Linear regression was used when a bivariate analysis involved a continuous variable (e.g., age).

The SPSS GLM routine was employed to combine variables into an overall predictive model. The continuous variables retained for this step in the analysis were entered as covariates. The GLM routine also was used to evaluate interaction terms. These analyses used the unique sums of squares option to estimate the contribution of a given predictor independent of the other predictors in the model. Following the GLM analysis, the SPSS regression routine was used to provide the final mathematical form of the model. The regression routine also provided regression diagnostics to identify outlier and influential data points (Belsley, Kuh, & Welsch, 1980; Stevens, 1984).

Results

Methods Analysis

All 3 methods variables predicted validity in the FE analyses:

A. *Test Type*. Fixed-time tests (r = .801) were more valid than fixed-distance tests (r = .726; χ^2 = 36.79, 1 *df*, $p < .001$).[9]
B. *SDVO$_{2max}$*. Validity increased with SDVO$_{2max}$ (r = .315; χ^2 = 46.85, 1 *df*, $p < .001$; z_{UF}' = .520 + .0672*SD).
C. *VO$_{2max}$ Protocol*. CTO validity (r = .704) was lower than that for other procedures (CTB, r = .783; IT, r = .783; Other, r = .786; χ^2 = 38.38, 3 *df*, $p < .001$).

The small differences in the average validities for the CTB, IT, and Other VO$_{2max}$ protocols suggested that those 3 procedures could be considered equivalent. Analysis of a dichotomous contrast between CTO and all other protocols confirmed that the dichotomy had virtually the same explanatory power as the 4-group classification (χ^2 = 38.35). The difference between the explanatory power of the dichotomy and that of the 4-group classification was trivial ($\Delta\chi^2$ = 0.03, 2 *df*, $p > .985$). Therefore, subsequent analyses treated VO$_{2max}$ Protocol as a dichotomy contrasting CTO with all other VO$_{2max}$ protocols.

Multivariate Methods Model

A multivariate model was developed to determine the minimum set of variables required to extract the explanatory power from the set of methods variables. The GLM routine was employed with Test Type and VO$_{2max}$ Protocol as FE group classification variables and SDVO$_{2max}$ as a continuous covariate. The analysis was limited to main effects because some Test Type-VO$_{2max}$ Protocol combinations were studied too infrequently to place much confidence in the detailed cell means that would be fitted if interactions were included in the model.

[9] These figures differ slightly from values reported in Vickers (2001b). The difference occurs because validity coefficients for each study were averaged in the present analyses. Each coefficient was treated separately in the earlier work.

The overall model was significant (χ^2 = 96.47, 3 *df*, *p* < .001). The largest effect was Test Type (χ^2 = 44.17, 1 *df*, *p* < .001). $SDVO_{2max}$ produced the next largest effect (χ^2 = 30.04, 1 *df*, *p* < .001). The effect of VO_{2max} Protocol was notably smaller than the others (χ^2 = 9.09, 1 *df*, *p* < .001). Expressed as a regression equation, the model was:

$$z_{UF(i)}' = 0.276T + .097V + .056S + .133 \quad \text{(Equation 1)}$$

where "T" indicates Test Type (1 = Fixed distance, 2 = Fixed time), "V" indicates the VO_{2max} Protocol (1= CTO, 2 = All Other), and "S" indicates $SDVO_{2max}$ in $ml \cdot kg^{-1} \cdot min^{-1}$.

Population Attributes

Population attributes might affect run test validity (Baumgartner & Jackson, 1982). If so, any attributes that affect validity would be bases for dividing the general population into subgroups with different validities. Meta-analysts would refer to the relevant attributes as moderators of run test validity (Hunter & Schmidt, 1990).

A. *Gender*. The difference between males (*r* = .726) and females (*r* = .755) was nonsignificant (χ^2 = 3.80, 1 *df*, *p* > .051, critical N = 1864)[10] in the simple bivariate analysis.[11] However, the average validity for males (*r* = .766) was lower than that for females (*r* = .804) after controlling for methods. This difference was statistically significant (χ^2 = 5.46, 1 *df*, *p* < .020), even though the absolute magnitude was small (critical N = 781).

B. *Age*. Validity was not related to age (*r* = -.015; χ^2 = 0.12, 1 *df*, *p* > .729; χ^2 = 0.27, 1 *df*, *p* > .603 controlling for methods).

C. *Experience*. Run tests were equally valid for endurance-trained athletes (*r* = .727) and untrained individuals (*r* = .733, χ^2 = 0.13, 1 *df*, *p* > .718; χ^2 = 0.70, 1 *df*, *p* > .402 controlling for methods).

D. *Fitness*. Average VO_{2max} was not related to validity (*r* = -.059; χ^2 = 1.67, 1 *df*, *p* > .196; χ^2 = 0.03, 1 *df*, p > .862 controlling for methods).

[10] Any difference in validity coefficients would be significant given a large enough sample size. Critical N is the smallest sample size that would make the observed difference significant at *p* < .05 (Hoelter, 1983). Effect sizes with a critical *N* > 200 can be considered too small to be important. Critical N is reported here to provide context for interpreting the observed difference.

[11] The critical Ns for differences between groups were computed using the estimated standard deviation $\sqrt{[1/(N_1 - 3) + 1/(N_2 - 3)]}$ (Hays, 1963, p. 532).

E. *Interaction effects*. Six 2-way interactions were analyzed.[12] The interactions of Athlete with Gender ($\chi^2 = 0.81$, 1 *df*, $p > .368$) and Athlete with Age ($\chi^2 = 1.53$, 1 *df*, $p > .216$), were not significant. The interactions of Gender with Age ($\chi^2 = 4.47$, 1 *df*, $p < .035$), Gender with VO_{2max} ($\chi^2 = 8.79$, 1 *df*, $p > .001$), Athlete with VO_{2max} ($\chi^2 = 14.47$, 1 *df*, $p < .001$), and Age with VO_{2max} ($\chi^2 = 8.66$, 1 *df*, $p < .001$) were statistically significant.

No interaction involving gender was significant controlling for methods (Gender-Age, $\chi^2 = 0.27$, 1 *df*, $p > .604$; Gender-VO_{2max}, $\chi^2 = 0.04$, 1 *df*, $p > .841$). The Athlete-VO_{2max} ($\chi^2 = 4.05$, 1 *df*, $p < .045$) and Age-VO_{2max} ($\chi^2 = 7.06$, 1 *df*, $p < .001$) interactions remained significant controlling for methods when analyzed separately.

The Athlete-VO_{2max} interaction was not significant ($\chi^2 = .003$, 1 *df*, $p > .956$) when the two remaining interactions were combined with methods in a single overall model. The Age-VO_{2max} interaction was significant in these analyses whether the Athlete-VO_{2max} interaction was in the model ($\chi^2 = 7.07$, 1 *df*, $p < .001$) or removed from it ($\chi^2 = 9.72$, 1 *df*, $p < .001$).

Random Effects Model

Hedges and Vevea's (1998) methods were employed to estimate an RE methods model. The χ^2 for the 3-variable RE model ($\chi^2 = 26.38$) was much smaller than that for the FE model ($\chi^2 = 97.82$).[13] VO_{2max} Protocol was not a significant predictor ($\chi^2 = 1.14$, 1 *df*, $p > .714$), so a revised RE model was computed with Test Type ($\chi^2 = 10.36$, 1 *df*, $p < .001$) and $SDVO_{2max}$ ($\chi^2 = 15.17$, 1 *df*, $p < .001$) as the only predictors. Expressed in regression form, the resulting 2-variable RE model was:

$$z_{UF}' = .262 + (.071*S) + (.227*T) \quad \text{(Equation 2)}$$

The 2-variable RE model was based on more data than the 3-variable model. Missing data limited the 3-variable model to 147 samples. VO_{2max} Protocol was the only missing data for 8 samples. The 2-variable RE model, therefore, was based on 155 samples. Both predictors contributed significantly to the explanatory power of the model

[12] The GLM procedure specified the model as main effects plus the interaction effects of interest. When the interaction involved a continuous variable, the interaction test was a test for parallelism of regression lines (Walker & Lev, 1953).

[13] The smaller χ^2 for the model was expected because this χ^2 is the sum of the squared standardized differences between predicted and observed values (Hays, 1963). Adding RE variance to sampling variance increases the estimated overall variance, so the standard deviation used in the standardization process increases. As a result, *z* values decrease, and the sum of the squared *z* values decreases. The latter sum is the overall model χ^2, so changing from an FE to an RE analysis is expected to decrease this overall indicator.

Table 1. Comparison of FE and RE Predictions

			Confidence Interval	
		Predicted	Lower	Upper
Test Type	Model	*r*	Bound	Bound
Fixed Distance	FE	.706	.690	.722
	RE	.724	.694	.750
Fixed Time	FE	.822	.797	.845
	RE	.815	.769	.853

Note. Estimates computed with $SDVO_{2max}$ = 6.00. Confidence intervals were computed using Hedges and Vevea's (1998) Equation 12. The Σw_i was based on the weights for the correlations for each test type, not for the full data set.

($SDVO_{2max}$, $\chi^2 = 20.80$, 1 *df*, $p < .001$; Test Type, $\chi^2 = 10.89$, 1 *df*, $p < .001$).

The effects of shifting from an FE model to an RE model cannot be determined by contrasting Equations 1 and 2. These equations represent different subsets of cases as well as different statistical models. Therefore, a 2-variable FE model was developed:

$$z_{UF}' = .194 + (.067*S) + (.284*T) \quad \text{(Equation 3)}$$

The effect of changing from an FE to an RE model can be determined by comparing the coefficients in Equation 3 to those in Equation 2. The RE intercept was higher and the RE coefficient for Test Type was lower. The RE coefficient for $SDVO_{2max}$ was virtually identical to the FE coefficient for this predictor.

The Age-VO_{2max} interaction was reevaluated after establishing the RE model. The interaction had been weak in the earlier analyses and might be statistically nonsignificant given the larger estimated sampling error in the RE model. This interaction was not significant in the RE model ($\chi^2 = 1.87$, 1 *df*, p > .171).[14]

Table 1 gives the validity estimates derived from the 2-variable FE and RE models (i.e., Equations 3 and 2, respectively). These estimates were obtained by computing z_{UF}', then reversing Fisher's *r*-to-*z* transformation. Relative to the FE model, the RE model produced a slightly higher estimated validity for fixed-distance tests and a

[14] Hedges and Vevea's (1998) RE modeling methods require iterative estimation of the weights. Two iterations were employed in the interaction model. The excess χ^2 at that point was only 0.51. Experience with this procedure in other analyses in this data set indicated that the impact of this residual excess on the weight variable would be so small that it would have no noticeable influence on the χ^2 values for the model if the analyses were extended beyond the second iteration.

slightly lower estimated validity for fixed-time tests. The choice of model had more effect on confidence intervals for the validity coefficients. The interval increased by .022 for fixed-distance tests and by .036 for fixed-time tests.

Outliers and Influential Data Points

Meta-analyses should include checks for the presence of outlier/influential data points (Hedges & Olkin, 1985; National Research Council, 1992). In this study, the search was initiated by fitting Equation 2 to the data as a regression model. The studentized-deleted residual, Cook's d, the centered leverage value, and DFFIT (cf., Belsley et al., 1980) were saved as indicators of the effect of each observation on the fit of the overall model to the data. Tukey's (1977) definition of an extreme data point was used to identify influential data points in box plots of these indicators.[15]

Eight validity coefficients were identified as influential data points. In 7 cases, the identification was based on an extreme value for Cook's d. In the 8th case, the identification was based the centered leverage value.

The initial identification was followed up by a detailed analysis of the nature of the influence of the extreme data points. The detailed analysis examined the impact of each data point on the regression coefficients for the model. The DFBETAs from the regression were used to indicate this effect (cf., Belsley et al., 1980). Six of the 7 coefficients with extreme Cook's d values were extremes in the distribution of DFBETA for $SDVO_{2max}$. Figure 1 (p. 16) shows that 5 of the 6 cases stood out from the bulk of the data because they combined a large $SDVO_{2max}$ with low validity. These 5 cases have been labeled as $SDVO_{2max}$ outliers.

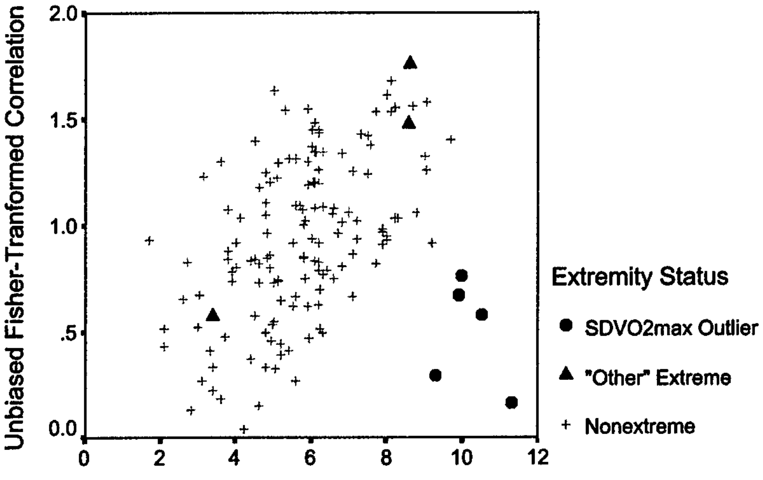
The fact that most of the influential data points shared a common characteristic made the exploration of this subset interesting. Thus, outlier effects were evaluated initially by eliminating just the 5 studies with extreme Cook's d values and extreme DFBETA for $SDVO_{2max}$.

Removing the 5 influential data points increased the explanatory power of the FE model ($\chi^2 = 95.42$ to $\chi^2 = 136.62$). This increase in variance explained was obtained even though the total χ^2 decreased ($\chi^2 = 472.52$ to $\chi^2 = 446.74$). Using the Tucker and Lewis (1973) index

An RE model was constructed with the 5 $SDVO_{2max}$ extremes deleted. As with the FE models, dropping the influential data points increased the overall explanatory power ($\chi^2 = 30.55$ to $\chi^2 = 62.93$).[16] The

[15] An extreme is a data point that is more than 3 times the interquartile range above the upper limit of that range or more than 3 times the range below its lower limit (Tukey, 1977).

[16] TLI values did not change. The RE computations produce a final $\chi^2 \approx$ degrees of freedom, so TLI $\approx$ 1.00 regardless of total χ^2.



Note. Extremity status based on outlier/influential data point statistics for 2-variable model. See text for details.

Figure 1. Plot of $z_{uf(i)}$ vs. $SDVO_{2max}$

correlation between $SDVO_{2max}$ and validity increased from r = .328 to r = .508. The correlation between Test Type and validity decreased from r = .231 to r = .224. The 2-variable RE model with outliers removed was

$$z_{UF}' = .059 + (.115*S) + .(197*T) \quad \text{(Equation 4)}$$

Comparing Equation 4 to Equation 2, removing the influential data points lowered the intercept by .203, raised the $SDVO_{2max}$ coefficient by .044, and lowered the Test Type coefficient by .030.

The other 3 coefficients initially identified as possible influential data points were reconsidered after fitting the RE model. First, their outlier status was determined with Equation 4 as the frame of reference. Only Johnson, Oliver, and Terry's (1979) sample of 100 males was extreme in this reassessment. This sample combined a large $SDVO_{2max}$ (8.59) with high validity (r = .90). However, the data point appeared to be influential primarily because of its large sample size rather than because this combination of validity coefficient and $SDVO_{2max}$ was exceptional. This inference is based on the evidence in Figure 1 showing that a number of other data points close to Johnson et al. (1979) were not influential. However, the Johnson et al. (1979) sample size was much larger than the average for this review.

Dropping the 3 additional possible influential data points had little effect on the model. Bivariate correlations of validity with the predictors were smaller ($SDVO_{2max}$, r = .508 to r = .471; Test Type, r = .224 to r = .190), and the FE model had less explanatory power (χ^2 = 136.62 to χ^2 = 96.49; TLI = .445 to TLI = .362). However, the regression coefficients for the model were largely unaffected. The coefficient for $SDVO_{2max}$ was .006 lower. The coefficient for Test Type was .004 lower. Thus, the primary effect of dropping these 3 cases was an increase in the intercept from .059 to .098.

On the whole, removing the 3 additional data points reduced the explanatory power of the model without changing its structure. Therefore, Equation 4 provided a model that combined wide coverage of the data with robustness to outlier/influential data points. Equation 4 was adopted as the final RE model.

Sensitivity Analyses

Meta-analytic findings should be checked for sensitivity to factors that might bias estimates (National Research Council, 1992). Sensitivity evaluations in this review produced the following:

A. *Search strategy.* The average correlation for the search based on VO_{2max} (r = .743) was not significantly different from that for the search based on threshold constructs (r = .715, χ^2 = 2.67, 1 *df*, p > .102).

B. *Publication status.* The average correlation for published studies (r = .753) was higher than that for unpublished studies (r = .703, χ^2 = 13.12, 1 *df*, p < .001). However, a single study had a major influence on this result. Dropping Fitzgerald et al. (1986) from the analysis reversed the difference (unpublished r = .794, χ^2 = 5.16, 1 *df*, p < .024).

C. *Publication date.* Correlations were slightly higher for earlier studies (z_{uf}' = 6.407-.0028*Year), but the trend was not significant (χ^2 = 1.94, 1 *df*, p > .163).

D. Choice of weights. The average Fisher-transformed correlation was r = .737 without weights compared with r = .740 with FE weights and r = .754 with RE weights.

E. Fisher-Transformation. Hunter and Schmidt's (1990) validity generalization approach to meta-analysis analyzes raw correlations rather than Fisher-transformed correlations. The unweighted average raw correlation was r = .681 for the 155 correlations in the final RE model. The weighted average was r = .707. The Fisher-transformed values were .056 and .047 higher, respectively (see paragraph D).

F. Underpowered studies. Kraemer et al. (1998) noted that studies with low statistical power can bias meta-analyses. Many coefficients in the present analysis were estimated in small samples, so lack of power was a concern. To evaluate the effect of sample size, analyses were repeated with just those correlations estimated in samples with $N \geq 15$. That sample size gives a power of .80 for a 1-tailed $p < .05$ significance test when $r = .70$. The sample size restriction eliminated 23 (15.3%) of the 150 samples that contributed to Equation 4, but had little effect on the model. The χ^2 dropped from 136.62 to 125.60. The regression coefficient decreased by .001 for $SDVO_{2max}$ while that for Test Type increased by .001. The intercept was .056 lower, and TLI decreased from .445 to .417. On the whole, the inclusion of underpowered studies had limited effects on the model.

G. *Large study effects*. Meta-analyses often include a few large samples among a number of much smaller samples (Osburn & Callender, 1992). In this review, Fitzgerald et al. (1986) provided 1,091 of the 5,757 observations for the study. The reported validity for the male sample in this study was lower than expected based on Equation 4 (χ^2 = 22.67), so Fitzgerald et al.'s (1986) results might have had a substantial impact on estimation of model components. The basic bivariate analyses were repeated with Fitzgerald removed along with the 5 $SDVO_{2max}$-validity outliers. The results in this reduced data set were basically the same as those in the original analyses. The methods model became $z_{UF}' = .101 + (.107*S) + (.206*T)$ (χ^2 = 114.19; TLI = .447). Thus, Fitzgerald et al. (1986) contributed to the misfit of the model, but did not distort the form of the model.

Run Test Precision

The results to this point suggested that endurance runs are valid estimators of aerobic capacity for all types of people. Given this observation, the study was extended to address another question. How precise are the VO_{2max} estimates derived from endurance run performance?

The question was answered by computing standard error of estimate (SEE) for each sample:

$$SEE_i = \sqrt{(1 - r_i^2)} * SDVO_{2max(i)}$$

The subscript indicates the i^{th} sample, and r is the raw (i.e., uncorrected, untransformed) validity coefficient for the sample.

Fixed-time and fixed-distance tests were considered separately. The higher average validity of fixed-time tests suggested that these tests would have a smaller SEE than fixed-distance tests.[17] Analysis supported this view:

[17] A smaller SEE would be expected if $SDVO_{2max}$ were the same for both types of test. Preliminary analyses suggested this condition was satisfied. The unweighted average $SDVO_{2max}$ was comparable for fixed-distance (SD = 5.86 ml·kg^{-1}·min^{-1}) and fixed-time (SD = 5.78 ml·kg^{-1}·min^{-1}) tests. Median values were even closer (5.82 ml·kg^{-1}·min^{-1} vs. 5.85 ml·kg^{-1}·min^{-1}). Boxplots identified several $SDVO_{2max}$ outliers among the fixed-distance tests, so the comparison was

A. *Fixed-distance tests*. The unweighted SEE was 3.85 $ml \cdot kg^{-1} \cdot min^{-1}$ (median = 3.78 $ml \cdot kg^{-1} \cdot min^{-1}$; range = 1.20 - 6.38 $ml \cdot kg^{-1} \cdot min^{-1}$) on the average. The weighted average was 4.20 $ml \cdot kg^{-1} \cdot min^{-1}$ (median = 4.50 $ml \cdot kg^{-1} \cdot min^{-1}$) using N - 1 as the weighting factor.

B. *Fixed-time tests*. The unweighted SEE was 3.34 $ml \cdot kg^{-1} \cdot min^{-1}$ (median = 3.34 $ml \cdot kg^{-1} \cdot min^{-1}$; range = 2.51 - 5.23 $ml \cdot kg^{-1} \cdot min^{-1}$) on the average. The weighted SEE was 3.41 $ml \cdot kg^{-1} \cdot min^{-1}$ (median = 3.37 $ml \cdot kg^{-1} \cdot min^{-1}$).

The SEE differences between tests type were larger than expected based on the average validity coefficients for the 2 test types. The difference might be explained by a stronger relationship between validity and $SDVO_{2max}$ for fixed-time tests than for fixed-distance tests. A multivariate analysis of covariance (MANCOVA) was conducted to test this possibility. Test Type was the group classification variable, and $SDVO_{2max}$ was a covariate. The statistical model included a test for the interaction between Test Type and $SDVO_{2max}$. The statistical test indicated that the slope of the regression of $z_{uf(i)}'$ on $SDVO_{2max}$ differed significantly between the test types for the full data set ($\chi^2 = 17.92$, 1 *df*, $p < .001$) and for the reduced data set with the 5 influential data points removed from the analysis ($\chi^2 = 9.59$, 1 *df*, $p < .001$).

Separate regression equations were computed for fixed-time and fixed-distance tests following the MANCOVA. This analysis corresponded to the common practice of decomposing statistically significant interaction effects into simple main effects. The analyses used $SDVO_{2max}$ to predict $z_{uf(i)}$. The results were:

A. *Fixed-distance tests*. The regression ($z_{uf(i)}' = .0738*SD + .447$) was statistically significant ($r = .332$; $\chi^2 = 37.44$, 1 *df*, $p < .001$). Residual variation was significant ($\chi^2 = 302.26$, 123 *df*, $p < .001$) and goodness of fit was low (TLI = .162.

B. *Fixed-time tests*. The regression ($z_{uf(i)}' = .153*SD + .221$) was significant ($r = .799$; $\chi^2 = 45.90$, 1 *df*, $p < .001$). Residual variation was not significant ($\chi^2 = 25.99$, 24 *df*, $p > .353$) and goodness of fit was high (TLI = .956).

The volume of missing data for $SDVO_{2max}$ in fixed-time studies was a significant concern when interpreting these findings. The observed differences might have been obtained because the 26 fixed-time tests with $SDVO_{2max}$ values were not representative of this test type. In fact, the average validity coefficient was significantly higher in studies

repeated for 124 fixed-distance samples with sample $SDVO_{2max}$ values between 2.0 $ml \cdot kg^{-1} \cdot min^{-1}$ and 10.0 $ml \cdot kg^{-1} \cdot min^{-1}$. Also, the weights were changed to N - 1 to reflect the weight actually used in computing the standard deviations. These changes increased the average $SDVO_{2max}$ for fixed-distance tests (mean = 6.11 $ml \cdot kg^{-1} \cdot min^{-1}$; median = 6.30 $ml \cdot kg^{-1} \cdot min^{-1}$) and fixed-time tests (mean = 6.24 $ml \cdot kg^{-1} \cdot min^{-1}$; median = 6.12 $ml \cdot kg^{-1} \cdot min^{-1}$).

with reported $SDVO_{2max}$ estimates (r_{with} = .828 vs. $r_{without}$ = .754, χ^2 = 11.70, 1 *df*, p < .001).[18]

Discussion

Two competing interpretations, one a substantial departure from Safrit et al.'s (1988) conclusions, could be proposed for the findings. If the difference between fixed-time and fixed-distance tests were emphasized, the run test validity estimate would increase from r = .66 to r = .82. With correction for attenuation due to measurement error, the estimated true validity would increase from r = .75 to r = .94. After correcting for restriction/enhancement of range effects, these validity estimates would apply to all populations and all situations. The overall conclusion would be that run test performance has a virtually perfect correspondence to VO_{2max} in all test settings.

The findings from the fixed-time tests should not be emphasized at this time. There is no logical theoretical basis for the difference between test types, and the existence of empirical differences is debatable. To begin with, studies that included both fixed-time and fixed-distance tests have either shown no difference (McNaughton et al., 1998, r = .87, n = 32; O'Gorman et al., 2000, r = .67, n = 15) or higher validity for fixed-distance tests (Knowlton & Gifford, 1972, r = .66 vs. r = .56, n = 20, z = 0.38, p > .703, 2-tailed; O'Donnell, Smith, O'Donnell, & Stacy, 1984, r = .83 vs. r = .72, n = 42, z = 1.75, p > .08, 2-tailed). Taken together, these studies suggest no difference between the different types of run test. Note, however, that the total sample size is modest. Even if validity does differ between the two types of test, the claim that the validity of fixed-time tests is more generalizable than that of fixed-distance tests may be based on a biased subset of the data. Only 68% (26 of 38) of available studies were included in the analysis that would be the basis for claiming complete generalizability of fixed-time validity. The possibility of bias in this subset of fixed-time studies is evident in the fact that the average validity coefficient was significantly lower in the 12 studies that were not included in the analysis because of missing $SDVO_{2max}$ values. The test type difference should be studied further, but it is not established well enough at this time to make it the basis for overall conclusions. The remainder of this discussion, therefore, treats all endurance run tests as a single group.

[18] Hunter and Schmidt's (1990) validity generalization (VG) approach was applied to the fixed-time data to further assess the generalizability of these tests. This analysis showed that $SDVO_{2max}$ accounted for 61.1% of the variation in validity coefficients for the fixed-time tests. The reliability of VO_{2max} tests and sampling variance accounted for another 35.0% of the variation. The total variation explained (96.1%) was well beyond Hunter and Schmidt's (1990) recommended 75% criterion for stopping the search for moderator variables.

Safrit et al.'s (1988) major findings have been updated in two regards. The estimated validity of endurance run tests increased from r = .66 to r = .75. One reason is that Safrit et al. (1988) included a significant proportion of runs that were too short to be included in this review. Comparing the weighted average of the raw correlations in this review (r = .71, p. 19) with Safrit et al.'s (1988) data, the criterion difference accounted for approximately .05 of the validity increase. This effect was predictable because earlier examinations of the relationship between test run distance/duration showed that more stringent criteria used here were needed to maximize validity (Vickers, 2001a, 2001b). The choice of meta-analytic procedure also contributed to the difference. Averaging raw correlations underestimates the population correlation; averaging Fisher-transformed correlations overestimates this value (Silver & Dunlap, 1987). Given these biases, the true population correlation value should be in the narrow range from r = .71 to r = .75.

This review also reinforced Safrit et al.'s (1988) assertion that the run test validity coefficient does not generalize across test situations. The analyses took a different statistical route to this conclusion by beginning with an FE model, then shifting to an RE model when the data showed significantly greater than chance residual variation. Translating the excess variation into an estimate of RE variance produced an RE interval that ranged from r = .52 to r = .87.[19] This interval could be narrowed by further research. The identification of sources of variation in validity that are not in the current model would narrow the range by reducing the RE variance. However, it may be unrealistic to expect substantial progress in this regard. The RE variance may reflect the cumulative variation arising from a number of causes and from interactions between those influences (Raudenbush, 1994, p. 302). If so, further study is not likely to change the range of probable values for run test validity.

The RE interval applies to all types of people. Age, gender, fitness, and running experience did not affect validity. Pairwise combinations of these attributes did not affect validity. The tests for moderating effects of attributes on validity should provide an accurate statistical basis for inference. The specific meta-analysis method should not affect the findings (Schmidt & Hunter, 1999). The significance tests should yield appropriate inferences given the clear evidence that an RE model is appropriate (Hedges & Vevea, 1998; Overton, 1998). However, the inferences drawn are subject to limitations of the data. Most samples consisted of people between 15

[19] The RE interval is based on the RE variance for the final model. The interval is 1.96*τ where τ is defined by Equation 10 in Hedges and Vevea (1998). The assumption that the distribution is reasonable if the variance is the sum of the effects of many small independent effects. The term "RE interval" was chosen over "confidence interval" and "credibility interval" because it is not clear that the interval of interest here corresponds to either term as they are used in the meta-analytic literature (e.g., Whitener, 1990).

and 40, so the evidence may understate age effects. The few studies that have sampled younger and older participants have not produced exceptional findings. However, additional studies of older and younger individuals would be useful as verification that the few available studies are representative of these populations. Also, tests for the effects of motivation are conspicuous in their absence. Physical fitness assessment experts have been concerned about this variable for more than 20 years. The information provided in the studies reviewed here did not provide any obvious basis for coding this characteristic.

Methods factors may be a promising area of investigation. The coverage of these factors was limited to VO_{2max} measurement procedures. Even those procedures were characterized in very general terms. Specific VO_{2max} Protocol characteristics, such as the amount of warm-up time, the method of setting initial work rate, the frequency and size of the increments in work rate, and so on, conceivably could affect the accuracy and reproducibility of the criterion measurements. If so, those factors would increase the observed validity coefficients by increasing criterion reliability. A detailed study of protocol attributes was beyond the scope of this review, but it could be a productive topic for future work.

VO_{2max} measurements are central to run test validation, but other methods factors also might account for some variation in validity coefficients. For example, time of day, running surface, and weather conditions might affect running performance. Performance also could be affected by whether the run was a test (e.g., for school or military) or a competitive race and whether it was completed individually or as part of a group. Although running is a common activity, practice at the specific run distance or time might help maximize performance. Knowledge that any of these factors affected validity would be a basis for better testing procedures. From the perspective of run test validity, RE variance would decrease and the RE interval would narrow. The narrower interval would apply to conditional validity determined by the specific test conditions and methods. This potential benefit shows that additional research effort may be worthwhile, but the results of that effort should not be taken for granted. In the final analysis, the RE variance may represent the confluence of a large number of small effects involving a wide range of situational factors and interactions among those factors. If the underlying sources of variance are complex, the uncertainty associated with a model such as the one developed here may be unavoidable.

Statistical artifacts affect run test validity. Samples with greater variation in VO_{2max} tended to produce higher validity coefficients. This effect was anticipated on the basis of the well-known effects of restriction/enhancement of range (Hunter & Schmidt, 1990). The usual formula for correcting for these effects was replaced by a regression analysis because differences in range can arise many ways. The formula may not be appropriate in all cases (Sackett & Yang, 2000). The regression technique used here made allowance for the sample-to-sample differences and provided a simple equation for estimating the effects of VO_{2max} variability on run test validity. The equation can be readily applied in any future studies to determine whether observed validity coefficients are consistent with the data summarized in this review.

This study did not allow for the effects of measurement error. Corrections for this artifact are a recommended element of Hunter and Schmidt's (1990) validity generalization approach to meta-analysis. This omission should be a concern, but it probably is not critical. Safrit et al. (1988) estimated that measurement error accounted for 17% of the variance in their review. The basic observations in this study would not change if differences in reliability coefficients accounted for the same proportion of observed variation. The 95% RE interval would narrow by .10 (i.e., .59 to .84). The sizable gain would still leave a wide range of plausible values. In fact, the modest effect of removing this RE variance underscores the challenge inherent in the current model. Future research would have to uncover several strong influences on validity to narrow the RE interval substantially.

This review provides a frame of reference for testing decisions. In applied testing, a situation-specific validity study is needed only if r = .52, the lower boundary of the 95% RE interval, is unacceptable. In this case, a validity study would be have 39-to-1 odds of demonstrating that validity was higher than this minimum in the test situation. If testing were undertaken merely to demonstrate that the validity coefficient was greater than zero, the lower bound of the RE interval could be used to estimate effect size when conducting a power analysis. Given this reference point, the minimum sample size required to achieve a power of β = .80 (p < .05, 1-tailed) would be N = 21.

The review findings also can help interpret validity studies. As a starting point, the evidence should foster recognition that sampling variance can yield small validity coefficients by chance even when the true validity is moderate. For example, if r = .52 and N = 21, the lower bound of the 95% confidence interval produced for the sample estimate of validity is r = .33. The results summarized here might motivate the researcher to replicate the findings rather than dismissing run tests as inadequate in the proposed testing situation. Closer consideration of the implications of the RE aspect of the model provided here would give further impetus to replication. The RE variance implies that the results of any study are affected by the unique configuration of causal influences present when the study is conducted. Under the usual interpretation of random effects, any replication of the original study will actually estimate a different validity coefficient than in the first study. Although key elements of the design can be reproduced, the full configuration of factors that comprise the study conditions cannot. Factors such as temperature, motivation of the subjects, and so on cannot be perfectly controlled by the investigator. Other factors, such as measurement error, are inherently random and beyond the control of the investigator. The net effect is that the result obtained in any attempt to replicate may vary widely from the earlier estimate because the details of the original situation are never perfectly reproduced. Finally, whether the study is replicated or not, the interpretation of the evidence cannot focus solely on the results obtained in the particular sample(s) under investigation. A large body of evidence indicates that validity is approximately r = .75. An extreme deviation in either direction should be viewed with skepticism and interpreted with appropriate caution.

The sensitivity analyses were encouraging. The basic validity estimate derived from the review was not sensitive to a number of potential sources of inaccuracy. Neither publication bias nor the fact that the typical study had a small sample size appears to have much effect on the findings. Search strategy affected the volume of evidence available, but not the basic validity findings. The one exceptionally large sample in the study had little effect on the final model parameters. The adoption of an RE model contributed to this observation because RE weights reduce the influence of large samples relative to small samples. The choice of a meta-analysis model did affect the findings, but those effects have been explored above. The choice may affect details of the findings, but it does not appear to affect the major empirical trends that have been interpreted in this discussion. Outliers were present, but had little effect. One reason is that only 3% of the validity coefficients were outliers. This figure is low compared with estimates of 10% to 20% in other research domains (Hedges, 1987; Hedges & Olkin, 1985). The outliers that were present affected the goodness of fit of the explanatory model, but not the parameters of the model.

Safrit et al. (1988) viewed the results of their review as a framework for future studies. This review has reinforced and extended that framework. Run tests are valid, but a single validity coefficient does not generalize across test situations. The estimated validity was higher (r = .75) in this review because of different inclusion criteria and analysis procedures. The generalizability issue has been clarified by providing an RE interval r = .52 to r = .87 for validity coefficients. The RE interval appears to apply regardless of the age, gender, fitness, or running experience of the population tested. The interval also is not affected by the method used to measure VO_{2max} in the laboratory. Statistical artifacts affect sample estimates of validity, but substantial RE variance would remain even if the effects of artifacts were completely eliminated. The reinforced and augmented framework provided here remains a framework. Issues raised in the interpretation of the findings suggest a number of lines of investigation that could elaborate on that framework. Further comparison of fixed-time and fixed-distance tests is the most pressing issue. If test type differences noted here were replicated in additional work, the uncertainties expressed above would be replaced by a single validity that applied to all test situations. However, details of the available evidence raise significant doubts that further investigation would reach this end point. The evidence reviewed here should be a useful point of reference for the design of future validity studies and, more important, the interpretation of the findings from those studies.

References

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Baumgartner, T. A., & Jackson, A. S. (1982). *Measurement for evaluation in physical education*. (2nd ed.). Dubuque, IA: Wm. C. Brown.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 357-382). New York: Russell Sage Foundation.

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: implications for situational specificity and validity generalization. *Personnel Psychology, 49*, 275-306.

Fitzgerald, P. I., Vogel, J. A., Daniels, W. L., Dziados, J. E., Teves, M. A., Mello, R. P., & Reich, P. J. (1986). The Body Composition Project: a summary report and descriptive data. Natick, MA: US Army Research Institute of Environmental Medicine.

Hays, W. L. (1963). *Statistics for Psychologists*. New York: Holt, Rinehart, Winston.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*(2), 443-455.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486-504.

Hoelter, J. W. (1983). The analysis of covariance structures: goodness-of-fit indices. *Sociological Methods and Research, 11*, 325-344.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis*. Newbury Park: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: cumulating research findings across studies*. Beverly Hills, CA: Sage Publications.

Iwaoka, K., Hatta, H., Atomi, Y., & Miyashita, M. (1988). Lactate, respiratory compensation threshold, and distance running performance in runners of both sexes. *International Journal of Sports Medicine, 9*, 306-309.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: assumptions, models, and data*. Beverly Hills, CA: Sage.

Johnson, D. J., Oliver, R. A., & Terry, J. W. (1979). Regression equation for prediction of performance in the twelve minute run walk test. *Journal of Sports Medicine, 19*, 165-170.

Katch, V., & Henry, F. M. (1972). Prediction of running performance from maximal oxygen debt and intake. *Medicine and Science in Sports, 4*(4), 187-191.

Knapik, J. (1989). The Army Physical Fitness Tests (APFT): a review of the literature. *Military Medicine, 154*, 326-329.

Knowlton, R. G., & Gifford, P. B. (1972). An evaluation of a fixed time and fixed distance task as performance measures to estimate aerobic capacity. *Journal of Sports Medicine and Physical Fitness, 12*(3), 163-170.

Kraemer, H. C., Gardner, C., Brooks, J. O., III, and Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23-31.

McNaughton, L., Hall, P., & Cooley, D. (1998). Validation of several methods of estimating maximal oxygen uptake in young men. *Perceptual and Motor Skills, 87,* 575-584.

National Research Council. (1992). *Combining information: statistical issues and opportunities for research.* Washington, DC: National Academy Press.

O'Donnell, C., Smith, D. A., O'Donnell, T. V., & Stacy, R. J. (1984). Physical fitness of New Zealand army personnel; correlation between field tests and direct laboratory assessments--anaerobic threshold and maximum O_2 uptake. *New Zealand Medical Journal, 97*(760), 476-479.

O'Gorman, D., Hunter, A., McDonnacha, C., & Kirwan, J. P. (2000). Validity of field tests for evaluating endurance capacity in competitive and international-level sports participants. *Journal of Strength and Conditioning Research, 14*(1), 62-67.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201-211.

Osburn, H. G., & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology,* 77, 115-122.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354-379.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 301-322). New York: Russell Sage Foundation.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Rosenthal, R., & DiMatteo, R. (2001). Meta-analysis: recent developments in quantitative methods for literature reviews. In S. T. Fiske, D. L. Schachter, & C. Zahn-Wexler (Eds.), *Annual Review of Psychology, Vol. 52 (pp. 59-82).* Palo Alto, CA: Annual Reviews, Inc.

Sackett, P. R., & Yang, H. (2000). Correction for restriction of range: an expanded typology. *Journal of Applied Psychology, 85,* 112-118.

Safrit, M. J., Hooper, L. M., Ehlert, S. A., Costa, M. G., & Patterson, P. (1988). The validity generalization of distance run tests. *Canadian Journal of Sports Science, 13*(4), 188-196.

Schmidt, F. L., & Hunter, J. E. (1999). Comparison of three meta-analysis methods revisited: an analysis of Johnson, Mullen, and Salas. *Journal of Applied Psychology, 84,* 144-148.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: should Fisher's *z* transformation be used? *Journal of Applied Psychology, 72,* 146-148.

SPSS, Inc. (1998a). *SPSS Base 8.0 Applications Guide.* Chicago: SPSS, Inc.

SPSS, Inc. (1998b). *SPSS Advanced Statistics.* Chicago: SPSS, Inc.

Steiger, J. H. (1980). Test for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245-251.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*(2), 334-344.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1-10.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Vickers, R. R., Jr. (2001a). *Running performance as an indicator of VO_{2max}: distance effects* (Technical Report No. 01-20). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr. (2001b). *Running performance as an indicator of VO_{2max}: a replication of distance effects* (Technical Report No. 01-24). San Diego, CA: Naval Health Research Center.

Walker, H. M., & Lev, J. (1953). *Statistical inference.* New York: Holt, Rinehart, and Winston.

Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74,* 259-264.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper and L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 42-55). New York: Russell Sage Foundation.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315-321.

Appendix A
Data Sources

Abe, D., Kazumasa, Y., Yamanobe, K., & Tamura, K. (1998). Assessment of middle-distance running performance in sub-elite young runners using energy cost of running. *European Journal of Applied Physiology, 77,* 320-325.

Acevedo, E. O., & Goldfarb, A. H. (1989). Increased training intensity effects on plasma lactate, ventilatory threshold, and endurance. *Medicine and Science in Sports and Exercise, 21*(5), 563-568.

Bar-Or, O., Zwiren, D., & Dotan, R. (1978). Correlations among aerobic fitness tests and VO_{2max} in men who vary in aerobic power. In R. J. Shepard & H. LaVallee (Eds.), *Physical Fitness Assessment* (pp. 356-362.). Springfield, IL: Charles C. Thomas.

Beckett, M. B., & Hodgdon, J. A. (1987). *Lifting and carrying capacities relative to physical fitness measures* (Technical Report No. 87-26). San Diego, CA: Naval Health Research Center.

Billat, V., Renoux, J. C., Pinoteau, J., Petit, B., & Koralsztein, J. P. (1994). Reproducibility of running time to exhaustion at VO_{2max} in subelite runners. *Medicine and Science in Sports and Exercise, 26*(2), 254-257.

Brandon, L. J., & Boileau, R. A. (1987). The contribution of selected variables to middle and long distance run performance. *Journal of Sports Medicine, 27,* 157-164.

Brandon, L. J., & Boileau, R. A. (1992). Influence of metabolic, mechanical and physique variables on middle distance running. *Journal of Sports Medicine and Physical Fitness, 32*(1), 1-9.

Bulbulian, R., Wilcox, A. R., & Darabos, B. L. (1986). Anaerobic contribution to distance running performance of trained cross-country athletes. *Medicine and Science in Sports and Exercise, 18*(1), 107-113.

Buono, M. J. (1987). *Validity of the 500 yard swim and 5 kilometer stationary cycle ride as indicators of aerobic fitness* (Technical Report No. 87-27). San Diego, CA: Naval Health Research Center.

Burke, E. J. (1976). Validity of selected laboratory and field tests of physical working capacity. *Research Quarterly, 47,* 95-103.

Burris, B. (1970). *Reliability and validity of the 12 minute run test for college women.* Paper presented at the AAHPER Convention, Seattle, WA.

Chen, J. A. (1991). *Selected physiological variables and distance running performance among non-elite, heterogeneous groups of male and female runners.* Unpublished master's thesis, Washington State University, Pullman, WA.

Cisar, C. J., Thorland, W. G., Johnson, G. O., & Housh, T. J. (1986). The effect of endurance training on metabolic responses and the prediction of distance running performance. *Journal of Sports Medicine, 26,* 234-240.

Claiborne, J. M. (1984). *Relationship of the anaerobic threshold and running performance in female recreational runners.* Unpublished doctoral dissertation, University of North Carolina at Greensboro.

Conley, D. S., Cureton, K. J., Hinson, B. T., Higbie, E. J., & Weyand, P. G. (1992). Validation of the 12-minute swim as a field test of peak aerobic power in young women. *Research Quarterly for Exercise and Sport, 63*(2), 153-161.

Conley, D. L., & Krahenbuhl, G. S. (1980). Running economy and distance running performance of highly trained athletes. *Medicine and Science in Sports and Exercise, 12*(5), 357-360.

Cooper, K. H. (1968). A means of assessing maximal oxygen intake. *Journal of the American Medical Association, 203*(3), 135-138.

Costill, D. L. (1967). The relationship between selected physiological variables and distance running performance. *Journal of Sports Medicine and Physical Fitness, 7*, 61-66.

Costill, D. L., Thomason, H., & Roberts, E. (1973). Fractional utilization of the aerobic capacity during distance running. *Medicine and Science in Sports, 5*(4), 248-252.

Crain, M. L. (1977). *The relationship of brachial pulse wave components and predicted VO_2 max to running performance.* Unpublished master's thesis, Northeast Missouri State University, Kirksville, MO.

Cunningham, L. N. (1990). Relationship of running economy, ventilatory threshold, and maximal oxygen consumption to running performance in high school females. *Research Quarterly for Exercise and Sport, 61*(4), 369-374.

Custer, S. J., & Chaloupka, E. C. (1977). Relationship between predicted maximal oxygen consumption and running performance of college females. *Research Quarterly, 48*, 47-50.

Davies, C. T. M., & Thompson, M. W. (1979). Aerobic performance of female marathon and male ultramarathon athletes. *European Journal of Applied Physiology, 41*, 233-245.

di Prampero, P. E., Atchou, G., Bruckner, J.-C., & Moia, C. (1986). The energetics of endurance running. *European Journal of Applied Physiology, 55*, 259-266.

di Prampero, P. E., Capelli, C., Pagliaro, P., Antonutto, G., Girardis, M., Zamparo, P., & Soule, R. G. (1986). Energetics of best performances in middle-distance running. *Journal of Applied Physiology, 74*(5), 2318-2324.

Dorociak, J. J. (1981). *Validity of running tests of 4, 8, and 12 minutes duration in estimating aerobic power for college women of different fitness levels.* Unpublished doctoral dissertation, Louisiana State University and Agricultural and Mechanical College, Baton Rouge, LA.

Duggan, A., & Tebbutt, S. D. (1990). Blood lactate at 12 km/h and vOBLA as predictors of run performance in non-endurance athletes. *International Journal of Sports Medicine, 11*(2), 111-115.

Evans, S. L., Davy, K. P., Stevenson, E. T., & Seals, D. R. (1995). Physiological determinants of 10-km performance in highly trained female runners of different ages. *Journal of Applied Physiology, 78*(5), 1931-1941.

Farrell, P. A., Wilmore, J. H., Coyle, E. F., Billing, J. E., & Costill, D. L. (1979). Plasma lactate accumulation and distance running performance. *Medicine and Science in Sports, 11*(4), 338-344.

Fay, L., Londeree, B. R., LaFontaine, T. P., & Volek, M. R. (1989). Physiological parameters related to distance running performance in female athletes. *Medicine and Science in Sports and Exercise, 21*(3), 319-324.

Fitzgerald, P. I., J. A. Vogel, et al. (1986). The Body Composition Project: a summary report and descriptive data. Natick, MA, US Army Research Institute of Environmental Medicine.

Florence, S.-L., & Weir, J. P. (1997). Relationship of critical velocity to marathon running performance. *European Journal of Applied Physiology, 75*, 274-278.

Foster, C. C., Jr. (1972). *Maximal aerobic power and the aerobic requirements of running in trained runners and trained non-runners.* Unpublished master's thesis, University of Texas at Austin.

Foster, C., Costill, D. L., Daniels, J. T., & Fink, W. J. (1978). Skeletal muscle enzyme activity, fiber composition, and VO_{2max} in relation to distance running performance. *European Journal of Applied Physiology, 39*, 73-80.

Forster, C., Daniels, J. T., & Yarbrough, R. A. (1977). Physiological and training correlates of marathon running performance. *Australian Journal of Sports Medicine, 9*, 58-61.

Getchell, L. H., Kirkendall, D., & Robbins, G. (1977). Prediction of maximal oxygen uptake in young adult women joggers. *Research Quarterly, 48*, 61-67.

Grant, J. A., Joseph, A. N., & Campagna, P. D. (1999). The prediction of VO_{2max}: a comparison of 7 indirect tests of aerobic power. *Journal of Strength and Conditioning Research, 13*(4), 346-352.

Grant, S., Corbett, K., Amjad, A. M., Wilson, J., & Aitchison, T. (1995). A comparison of methods of predicting maximum oxygen uptake. *British Journal of Sports Medicine, 29*, 147-152.

Gutin, B., Fogle, R. K., & Stewart, K. (1976). Relationship among submaximal heart rate, aerobic power, and running performance in children. *Research Quarterly, 47*, 536-540.

Hagan, R. D., Smith, M. G., & Gettman, L. R. (1981). Marathon performance in relation to maximal aerobic power and training indices. *Medicine and Science in Sports and Exercise, 13*(3), 185-189.

Hagan, R. D., Upton, S. J., Duncan, J. J., & Gettman, L. R. (1987). Marathon performance in relation to maximal aerobic power and training indices in female distance runners. *British Journal of Sports Medicine, 21*(1), 3-7.

Harrison, M. H., Bruce, D. L., Brown, G. A., & Cochrane, L. A. (1980). A comparison of some indirect methods for predicting maximal oxygen uptake. *Aviation, Space, and Environmental Medicine, 51*(10), 1128-1133.

Haverty, M., Kenney, W. L., & Hodgson, J. L. (1988). Lactate and gas exchange responses to incremental and steady state running. *British Journal of Sports Medicine, 22*(2), 51-54.

Hazard, A. A. (1982). *The effects of endurance training at 2,440 m altitude on maximal oxygen uptake at altitude and sea level in young male and female middle distance runners.* Unpublished master's thesis, San Diego State University, CA.

Hodgdon, J. A., Vickers, R. R., Jr., & Bennett, B. L. (1983). *Impact of physiological and psychological factors on performance in a middle-distance run* (Technical Report No. 80-30). Naval Health Research Center.

Houmard, J. A., Costill, D. L., Mitchell, J. B., Park, S. H., & Chenier, T. C. (1991). The role of anaerobic ability in middle distance running performance. *European Journal of Applied Physiology and Occupational Physiology, 62*(1), 40-3.

Houmard, J. A., Craig, M. W., O'Brien, K. F., Smith, L. L., Israel, R. G., & Wheeler, W. S. (1991). Peak running velocity, submaximal energy expenditure, VO_{2max}, and 8 km distance running performance. *Journal of Sports Medicine and Physical Fitness, 31*(3), 345-350.

Huhn, R. R. (1975). *The reliability, validity, and predictability of twelve- and fifteen-minute field tests in relation to laboratory maximal oxygen uptake tests*. Unpublished master's thesis, San Diego State University, CA.

Jackson, A., der Weduwe, K., Schick, R., & Sanchez, R. (1990). An analysis of the validity of the three-mile run as a field test of aerobic capacity in college males. *Research Quarterly, 61*(3), 233-237.

Jackson, A. S., & Coleman, A. E. (1976). Validation of distance run tests for elementary school children. *Research Quarterly, 47*, 86-94.

Johnson, D. J., Oliver, R. A., & Terry, J. W. (1979). Regression equation for prediction of performance in the twelve minute run walk test. *Journal of Sports Medicine, 19*, 165-170.

Katch, F. I., McArdle, W. D., Czula, R., & Pechar, G. S. (1973). Maximal oxygen intake, endurance running performance, and body composition in college women. *Research Quarterly, 44*(3), 301-312.

Katch, V.I. (1970). The role of maximal oxygen intake in endurance performance. Paper presented at the AAHPER Convention, Seattle, WA. (Cited in Safrit et al., 1988).

Katch, V., & Henry, F. M. (1972). Prediction of running performance from maximal oxygen debt and intake. *Medicine and Science in Sports, 4*(4), 1887-1891.

Kearney, J. T., & Byrnes, W. C. (1974). Relationship between running performance and predicted maximum oxygen uptake among divergent ability groups. *Research Quarterly, 45*(1), 9-15.

Kitagawa, K., Miyashita, M., & Yamamoto, K. (1977). Maximal oxygen uptake, body composition, and running performance in young Japanese adults of both sexes. *Japanese Journal of Physical Education, 21*(6), 335-340.

Kitagawa, K., Yamamoto, K., & Miyashita, M. (1978). Maximal oxygen uptake, body composition and running performance in Japanese young adults of both sexes. In F. Landry & W. A. R. Orban (Eds.), *Exercise physiology: fitness and performance capacity studies* (pp. 553-561). Miami, FL: Symposia Specialists, Inc.

Knowlton, R. G., & Gifford, P. B. (1972). An evaluation of a fixed time and fixed distance task as performance measures to estimate aerobic capacity. *Journal of Sports Medicine and Physical Fitness, 12*(3), 163-170.

Kumagai, S., Tanaka, K., Matsuura, Y., Matsuzaka, A., Hiraboba, K., & Asano, K. (1982). Relationships of the anaerobic threshold with the 5 km, 10 km, and 10 mile races. *European Journal of Applied Physiology, 49*, 13-23.

Lacour, J. R., Padilla-Magunacelaya, S., Barthelemy, J. C., & Dormois, D. (1990). The energetics of middle-distance running. *European Journal of Applied Physiology, 60*, 38-43.

Lambert, G. P. (1990). *The relationship between physiological measurements and cross-country running performance.* Unpublished master's thesis, Ball State University, Muncie, IN.

Laukkanen, R., Oja, P., Pasanen, M., & Vuori, I. (1992). Validity of a two kilmetre walking test for estimating maximal aerobic power in overweight adults. *International Journal of Obesity, 16*, 263-268.

Leach, D. A. (1983). *The measurement of cardio-respiratory endurance and the Standard Evaluation for Army personnel in the 40-45 age category*. Carlisle, PA: U.S. Army War College, Study Project. (Cited in Knapik, 1989).

Lehmann, M., Berg, A., Kapp, R., Wessinhage, T., & Keul, J. (1983). Correlations between laboratory testing and distance running performance in marathoners of similar performance ability. *International Journal of Sports Medicine, 4*(4), 226-230.

Loftin, M., Zingraf, S., Warren, B., Jones, C. J., Brandon, J. E., & Harsha, D. (1986). Influence of physiological function and perceptual effort on 1.5 miles performance in college women. *Journal of Sports Medicine, 26*, 214-218.

MacNaughton, L., Croft, R., Pennicott, J., & Long, T. (1990). The 5 and 15 minute runs as predictors of aerobic capacity in high school students. *Journal of Sports Medicine and Physical Fitness, 30*, 24-28.

Mahon, A. D., Del Corral, P., How, C. A., Duncan, G. E., & Ray, M. L. (1996). Physiological correlates of 3-kilometer running performance in male children. *International Journal of Sports Medicine, 17*(8), 580-584.

Maksud, M. G., Cannistra, C., & Dublinski, D. (1976). Energy expenditure and VO_{2max} of female athletes during treadmill exercise. *Research Quarterly, 47*, 692-697.

Maksud, M. G., & Coutts, K. D. (1971). Application of the Cooper twelve-minute run-walk test to young males. *Research Quarterly, 42*(1), 54-59.

Massicotte, D. R., Gauthier, R., & Markon, P. (1985). Prediction of VO_{2max} from the running performance in children aged 10-17 years. *Journal of Sports Medicine, 25*, 10-17.

Matsui, H., Miyashita, M., Miura, M., Kobayoshi, K., Hoshikawa, T., & Kamei, S. (1972). Maximum oxygen intake and its relationship to body weight of Japanese adolescents. *Medicine and Science in Sports, 4*(1), 29-32.

Maughan, R. J., & Leiper, J. B. (1983). Aerobic capacity and fractional utilisation of aerobic capacity in elite and non-elite male and female marathon runners. *European Journal of Applied Physiology, 52*, 80-87.

Mayhew, J. L., & Andrew, J. (1975). Assessment of running performance in college males from aerobic capacity percentage utilization coefficients. *Journal of Sports Medicine, 15,* 342-345.

McCutcheon, M. C., Stichar, S. A., Giese, M. D., & Nagle, F. J. (1990). A further analysis of the 12-minute run prediction of maximal aerobic power. *Research Quarterly for Exercise and Sport, 61*(3), 280-283.

McNaughton, L., Hall, P., & Cooley, D. (1998). Validation of several methods of estimating maximal oxygen uptake in young men. *Perceptual and Motor Skills, 87,* 575-584.

Mello, R. P., Murphy, M. M., & Vogel, J. A. (1984). *Relationship between the Army two mile run test and maximal oxygen uptake* (Technical Report No. T3/85). City, ST: U.S. Army Research Institute of Environmental Medicine.

Morgan, D. W., Baldini, F. D., Martin, P. E., & Kohrt, W. M. (1989). Ten kilometer performance and predicted velocity at VO_{2max} among well-trained male runners. *Medicine and Science in Sports and Exercise, 21*(1), 78-83.

Morgan, D. W., & Daniels, J. T. (1994). Relationship between VO_{2max} and the aerobic demand of running in elite distance runners. *International Journal of Sports Medicine, 15*(7), 426-429.

Morlang, C. (1988). *The effects of phosphate loading on red blood cell 2,3-DPG and endurance running performance.* Unpublished master's thesis San Diego State University, CA.

Mosenthal, T. M. (1988). *Correlations of laboratory tests to distance running performance during a cross-country track season.* Unpublished master's thesis, St. Cloud State University, MN.

Myles, W. S., Brown, T. E., & Pope, J. I. (1980). A reassessment of a running test as a measure of cardiorespiratory fitness. *Ergonomics, 23*(6), 543-547.

Myles, W. S., & Toft, R. J. (1982). A cycle ergometer test of maximal aerobic power. *European Journal of Applied Physiology and Occupational Physiology, 49*(1), 121-129.

Noakes, T. D., Myburgh, K. H., & Schall, R. (1990). Peak treadmill running velocity during the VO_{2max} test predicts running performance. *Journal of Sports Sciences, 8,* 35-45.

O'Donnell, C., Smith, D. A., O'Donnell, T. V., & Stacy, R. J. (1984). Physical fitness of New Zealand army personnel; correlation between field tests and direct laboratory assessments--anaerobic threshold and maximum O_2 uptake. *New Zealand Medical Journal, 97*(760), 476-479.

O'Gorman, D., Hunter, A., McDonnacha, C., & Kirwan, J. P. (2000). Validity of field tests for evaluating endurance capacity in competitive and international-level sports participants. *Journal of Strength and Conditioning Research, 14*(1), 62-67.

Oja, P., Laukkanen, R., Pasanen, M., Tyry, T., & Vuori, I. (1991). A 2-km walking test for assessing the cardiorespiratory fitness of healthy adults. *International Journal of Sports Medicine, 12*(4), 356-362.

Paavolinen, L. M., Nummela, A. T., & Rusko, H. K. (1999). Neuromuscular characteristics and muscle power as determinants of

5-km running performance. *Medicine and Science in Sports and Exercise, 31*(1), 124-130.

Padilla, S., Bourdin, M., Barthelemy, J. C., & Lacour, J. R. (1992). Physiological correlates of middle-distance running performance. *European Journal of Applied Physiology, 65*, 561-566.

Palgi, Y., Gutin, B., Young, J., & Alehandro, D. (1984). Physiologic and anthropometric factors underlying endurance performance in children. *International Journal of Sports Medicine, 5*(2), 67-73.

Peronnet, F., Thibault, G., Rhodes, E. C., & McKenzie, D. C. (1987). Correlation between ventilatory threshold and endurance capability in marathon runners. *Medicine and Science in Sports and Exercise, 19*(6), 610-615.

Porcari, J., Freedson, P., Ward, A., Rippe, J., Wilkie, S., Kline, G., Keller, B., & Hsieh, S. (1987). Prediction of VO_{2max} using the ACSM VO_2 prediction for running. *Medicine and Science in Sports and Exercise, 19*, S29.

Powers, S. K., Dodd, S., Deason, R., Byrd, R., & McKnight, T. (1983). Ventilatory threshold, running economy and distance running performance of trained athletes. *Research Quarterly for Exercise and Sport, 54*(2), 179-182.

Priest, J. W., & Hagan, R. D. (1987). The effects of maximum steady state pace training on running performance. *British Journal of Sports Medicine, 21*(1), 18-21.

Ramsbottom, R., Nute, M. G. L., & Williams, C. (1987). Determinants of five kilometre running performance in active men and women. *British Journal of Sports Medicine, 21*(2), 9-13.

Ramsbottom, R., Williams, C., Boobis, L., & Freeman, W. (1989). Aerobic fitness and running performance of male and female recreational runners. *Journal of Sports Sciences, 7*, 9-20.

Ramsbottom, R., Williams, C., Fleming, N., & Nute, M. L. G. (1989). Training induced physiological and metabolic changes associated with improvements in running performance. *British Journal of Sports Medicine, 23*(3), 171-176.

Rasch, P. J. (1974). Maximal oxygen intake as a predictor of performance in running events. *Journal of Sports Medicine, 14*, 32-39.

Rasch, P. J., & Wilson, D. (1964). The correlation of selected laboratory tests of physical fitness with military endurance. *Military Medicine*, 256-258.

Ribisl, P. M., & Kachadorian, W. A. (1969). Maximal oxygen intake prediction in young and middle-aged males. *Journal of Sports Medicine, 9*, 17-22.

Rudzki, S. J. (1989). Weight-load marching as a method of conditioning Australian army recruits. *Military Medicine, 154*(4), 201-205.

Salzer, D. W. (1996). *The relationship of strength to endurance while exercising in a chemical warfare uniform in the heat*. Unpublished master's thesis, San Diego State University, CA.

Schrader, T. A. (1982). *Fluid ingestion and long-distance running.* Unpublished master's thesis, Arizona State University, Tempe, AZ.

Scrimgeour, A. G., Noakes, T. D., Adams, B., & Myburgh, K. (1986). The influence of weekly training distance on fractional utilization of maximum aerobic capacity in marathon and ultramarathon runners. *European Journal of Applied Physiology, 55*, 202-209.

Seljevold, P. J. (1989). *Prediction of running performance from selected variables measured during bicycle ergometry.* Unpublished master's thesis, St. Cloud State University, MN.

Shaver, L. G. (1975). Maximum aerobic power and anaerobic work capacity prediction from various running performances of untrained college men. *Journal of Sports Medicine, 15*, 147-150.

Sidney, K. H., & Shepard, R. J. (1977). Maximum and submaximum exercise tests in men and women in the seventh, eighth, and ninth decades of life. *Journal of Applied Physiology: Respiratory, Environmental, and Exercise Physiology, 43*(2), 280-287.

Sjodin, B., & Svedenhag, J. (1985). Applied physiology of marathon running. *Sports Medicine, 2*, 83-99.

Sparling, P. B., & Cureton, K. J. (1983). Biological determinants of the sex difference in 12-min run performance. *Medicine and Science in Sports and Exercise, 15*(3), 218-223.

Takeshima, N., & Tanaka, K. (1995). Prediction of endurance running performance for middle-aged and older runners. *British Journal of Sports Medicine, 29*, 20-23.

Tanaka, K., & Matsuura, Y. (1982). A multivariate analysis of the role of certain anthropometric and physiological attributes in distance running. *Annals of Human Biology, 9*(5), 473-482.

Tanaka, K., & Matsuura, Y. (1984). Marathon performance, anaerobic threshold, and onset of blood lactate accumulation. *Journal of Applied Physiology, 57*(3), 640-643.

Tanaka, K., Matsuura, Y., Matsuzaka, A., Hirakoba, K., Kumagai, S., Sun, S. O., & Asano, K. (1984). A longitudinal assessment of anaerobic threshold and distance-running performance. *Medicine and Science in Sports and Exercise, 16*(3), 278-282.

Tanaka, K., Takeshima, N., Kato, T., Niihata, S., & Ueda, K. (1990). Critical determinants of endurance performance in middle-aged and elderly endurance runners with heterogeneous training habits. *European Journal of Applied Physiology, 59*, 443-449.

Tanaka, K., Watanabe, H., Konishi, T., Mitsuzono, R., Sumida, S., Tanaka, S., Fukuda, T., & Nakadomo, F. (1986). Longitudinal associations between anaerobic threshold and distance running performance. *European Journal of Applied Physiology, 55*, 248-252.

Thiart, B. F., Blaauw, J. H., & van Rensburg, J. P. (1978). Endurance training and the VO2 max with special reference to validity of the Astrand-Rhyming nomogram and the Cooper 12-minute run as indirect tests for maximal oxygen uptake. In F. Landry & W. A. R. Orban (Eds.), *Exercise physiology: fitness and performance capacity studies* (pp. 609-614). Miami, FL: Symposia Specialists, Inc.

Tokmakidis, S. P., & Leger, L. A. (1992). Comparison of mathematically determined blood lactate and heart rate "threshold" points and relationship with performance. *European Journal of Applied Physiology, 64*, 309-317.

Tokmakidis, S. P., Leger, L. A., & Pilianidis, T. C. (1998). Failure to obtain a unique threshold on the blood lactate concentration curve during exercise. *European Journal of Applied Physiology and Occupational Physiology, 77*(4), 333-342.

Trone, D. W. (1989). *Predicting 10 km run performance time from physiological measurements*. Unpublished master's thesis, San Diego State University, CA.

Unnithan, V. B., Timmons, J. A., Paton, J. Y., & Rowland, T. W. (1995). Physiological correlates to running performance in pre-pubertal distance runners. *International Journal of Sports Medicine, 16*(8), 528-533.

Walters, S. C. (1983). *The physiological effects of a twenty week distance running program on teenage girls correlated with performance.* Unpublished master's thesis, Arizona State University, Tempe, AZ.

Wannamaker, G. S. (1970). A study of the validity and reliability of 12-minute run under selected motivational conditions. *American Corrective Therapy Journal, 24*(3), 69-72.

Ward, A., Wilkie, S., O'Hanley, S., Trask, C., Kallmes, D., Kleinerman, J., Crawford, B., Freedson, P., & Rippe, J. (1987). Estimation of VO_{2max} in overweight females [Abstract]. *Medicine and Science in Sports and Exercise, 19,* S29.

Weltman, J., Seip, R., Levine, S., Snead, D., Rogol, A., & Weltman, A. (1989). Prediction of lactate threshold and fixed blood lactate concentrations from 3200-m time trial running performance in untrained females. *International Journal of Sports Medicine, 10*(3), 207-211.

Weyand, P. G., Cureton, K. J., Conley, D. S., Sloniger, M. A., & Liu, Y. L. (1994). Peak oxygen deficit predicts sprint and middle-distance track performance. *Medicine and Science in Sports and Exercise, 26*(9), 1174-1180.

Wiley, J. F., & Shaver, L. G. (1972). Prediction of maximum oxygen intake from running performance of untrained young men. *Research Quarterly, 43*(1), 89-93.

Williams, K. R., & Cavanagh, P. R. (1987). Relationship between distance running mechanics, running economy, and performance. *Journal of Applied Physiology, 63*(3), 1236-1245.

Wiswell, R. A., Jaque, S. v., Marcell, T. J., Hawkins, S. A., Tarpenning, K. M., Constantino, N., & Hyslop, D. M. (2000). Maximal aerobic power, lactate threshold, and running performance in master athletes. *Medicine and Science in Sports and Exercise, 32*(6), 1165-1170.

Wyndham, C. H., Strydom, N. B., van Graan, C. H., van Rensburg, A. J., Rogers, G. G., Greyson, J. S., & van der Walt, W. H. (1971). Walk or jog for health: II. Estimating the maximum aerobic capacity for exercise. *South African Medical Journal, 45,* 53-57.

Yoshida, T., Chida, M., Ichioka, M., & Suda, Y. (1987). Blood lactate parameters related to aerobic capacity and endurance performance. *European Journal of Applied Physiology, 56,* 7-11.

Yoshida, T., Ishiko, T., & Muraoka, I. (1983). Cardiorespiratory functions in children with high and low performances in endurance running. *European Journal of Applied Physiology, 51,* 313-319.

Yoshida, T., Udo, M., Iwai, K., Chida, M., Ichioka, M., Nakadomo, F., & Yamaguchi, T. (1990). Significance of the contribution of aerobic and anaerobic components to several distance running performances in female athletes. *European Journal of Applied Physiology, 60*, 249-253.

Zacharogiannis, E., & Farrally, M. (1993). Ventilatory threshold, heart rate deflection point and middle distance running performance. *Journal of Sports Medicine and Physical Fitness, 33*(4), 337-347.

Zwiren, L. D., Freedson, P. S., Ward, A., Wilke, S., & Rippe, J. M. (1991). Estimation of VO_{2max}: a comparative analysis of five exercise tests. *Research Quarterly for Exercise and Sport, 62*(1), 73-78.

REPORT DOCUMENTATION PAGE

1. Report Date (DD MM YY)
26 Sep 2002

2. Report Type
Technical Interim

3. DATES COVERED (from - to)
Sep 2000 - Aug 2002

4. TITLE AND SUBTITLE
Modeling Run Test Validity: A Meta-Analytic Approach

5a. Contract Number: US Army Reimb
5b. Grant Number:
5c. Program Element:
5d. Project Number:
5e. Task Number:
5f. Work Unit Number: 60109

6. AUTHORS
Ross R. Vickers, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

9. PERFORMING ORGANIZATION REPORT NUMBER
Report No. 02-27

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

MED-02
2600 E St NW
Washington DC 20372-5300

10. Sponsor/Monitor's Acronyms(s)

11. Sponsor/Monitor's Report Number(s)

12 DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT (maximum 200 words)

Previous summaries of the research evidence have shown that run tests are valid indicators of VO_{2max}, but analysis also indicated that validity differed from one test situation to another. This study utilized data from 166 samples (N = 5,757) to test the general hypothesis that differences in testing methods could account for the cross-situational variation in validity. Only runs $\geq$2 km or $\geq$12 min were included. These criteria restricted attention to tests with maximal validity. The estimated average validity (r = .75). Validity was higher for fixed-time runs than for fixed-distance runs and in samples with greater variability in VO_{2max}. This difference must be interpreted cautiously because studies that directly compared these 2 types of run test have found little or no difference. Validity was not related to the age, gender, fitness, or running experience of the population tested or to the method used to measure VO_{2max}. A random-effects model estimated the 95% credibility interval for the validity of run tests at r = .52 to r = .84. The evidence was consistent with the view that some methods factors affect run test validity, but tests are equally valid for different types of people. This summary provides a point of departure for the design and interpretation of future run test validation studies.

15. SUBJECT TERMS
physical fitness, aerobic fitness, run tests, meta-analysis, validity

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	UNCL	41	Commanding Officer
UNCL	UNCL	UNCL			**19b. TELEPHONE NUMBER (INCLUDING AREA CODE)** COMM/DSN: (619) 553-8429